Phonology **26** (2009). Supplementary materials
doi: 10.1017/S0952675709990145

On the role of locality in learning stress patterns

Jeffrey Heinz
University of Delaware

Supplementary materials

Appendix A: Result of the neighbourhood learning study

The tables below describe the stress patterns in the database and the results of the learners discussed in the paper. The best descriptions of the stress patterns would be a description of the finite-state machines themselves, such as regular expressions, but these are unfortunately often too long to include in this appendix and difficult to read. Instead I rely on Bailey's (1995) Syllable Priority Code and some extensions I develop, which provides a short and readable – though imperfect – way of describing the patterns. The machines themselves are available (August 2009) at http://phonology.cogsci.udel.edu/dbs/stress/.

The tables below should be interpreted as follows. The 'name' column provides the name of a language in the typology which exemplifies the pattern. Information in parentheses following the language name refers either to the domain of the pattern (e.g. nouns) or to the researchers who described the pattern (full references which do not occur in the paper can be found below). The 'primary stress' and 'secondary stress' columns provide a Syllable Priority Code describing the stress, discussed in some detail below. The 'note' column indicates whether or not there are any phonotactic restrictions (which the sample obeys) or other relevant information. In particular, X indicates that the stress pattern is not tail-canonically neighbourhood-distinct and Y that it is not head-canonically neighbourhood-distinct. Thus, the absence of X (or Y) indicates tail (or head) canonical neighbourhood-distinctness. In the 'FL', 'BL' and 'FBL' columns (Forward Learner, Backward Learner and Forward Backward Learner respectively), the number indicates which forms were necessary for convergence. Specifically, *n* means the learner succeeded in learning the pattern given a sample of words consisting of every word which obeyed the pattern which was between 1 and *n* syllables in length. Failure of the learner is marked with *.

The Syllable Priority Code (SPC) in the 'primary stress' column was developed by Bailey (1995) as a shorthand for indicating primary stress assignment rules. The last character of the SPC (L or R) indicates which edge of the word counting begins from. The initial syllable is thus designated 1L, the peninitial 2L, the penultimate 2R and the final syllable 1R. Thus the simplest SPC codes, such as

1L for Afrikaans (Table I), simply indicate that main stress falls on the initial syllable.

Generally, more complex SPCs can be read as a series of if-then-else statements. Slashes indicate a quantity-sensitive rule, with rules governing heavier syllables occurring to the left of the slash. Thus the SPC 12/2L for Maidu (Table IV) unpacks as follows: if the initial syllable is heavy, it receives stress, else if the peninitial syllable is heavy, it receives stress, else stress falls on the peninitial syllable. This is redundant in this case, but I follow Bailey's (1995) conventions as far as possible. When a syllable position is suffixed with @s, primary stress occurs on that position only if it carries secondary stress; e.g. 12@sL for Malakmalak (Table II) means that primary stress falls on the initial syllable if it has secondary stress, else on the peninitial syllable if has secondary stress.

Unbounded patterns, where the stress can fall any distance from the word edge, use the 12..89 construct. For example, the SPC for Amele, 12..89/1L (Table III), unpacks as follows: if the first syllable counting from the left is heavy, it receives primary stress, else if the second syllable counting from the left is heavy, it receives primary stress … otherwise (if there are no heavy syllables) the first syllable counting from the left receives primary stress. Since words are unbounded in length, Bailey (1995) uses ..89 to indicate 'and so on'. Thus 89 does not literally mean the eighth or ninth syllable. Rather, 9 means the furthest syllable from the relevant edge and 8 the next-to-furthest syllable from the relevant edge and so on. See Bailey (1995) for more details.

If an SPC is followed by (n+), the code only applies to words that have at least n syllables. Likewise, an SPC followed by (n−) means the code only applies to words that have at most n syllables.

The 'secondary stress' column contains extensions I made to the SPC in order to describe secondary stress patterns. 'None' means that no secondary stress is present. 'Not included' means that source material reports secondary stresses, but either (i) the source material did not describe it, usually because it was deemed too complex, or (ii) the source material did describe it, but the pattern was either unclear or too complicated for me to incorporate into the study due to lack of time.

I indicate secondary stress patterns that can be described iteratively with the prefix 'i'. The prefix i2 means a syllable two syllables away from a stress (in either direction) receives a stress. The referent stress is indicated with a SPC suffixed with the @ symbol. Thus i2@1L for Bagandji (Table II) indicates that stress falls on the initial syllable and every other syllable, i.e. odd syllables from the left, whereas i2@2R for Anejom (Table II) indicates that secondary stresses fall on even syllables from the right. @m means that the first stress upon which the iterative procedure is based is the position of main stress. @mL means the iterations proceed only leftwards from main stress. Similarly, @mR means the iterations proceed only rightwards from main stress.

When the secondary stress rules are quantity-sensitive, I use H, L and X to designate heavy, light and either heavy or light syllables respectively. Thus a typical trochaic pattern is designated i('H,'LL) and a typical iambic pattern i(H',LX'). This description is ambiguous for certain strings, but the machines themselves are unambiguous – this is where readability is preferred over precision. If the iterative procedure begins from the word edge (as opposed to from a particular position), I forego the connective @ and use just the suffix L or R to

indicate whether the pattern proceeds from the left or right edge respectively. Thus i('H,'LL)R for Ancient Greek (Table V) means trochees are iteratively constructed from the right word edge.

Whenever only heavy syllables bear secondary stress, I indicate this with H. Sometimes it is necessary to mention explicitly that secondary stress only precedes main stress (as in cases describable with foot extrametricality), in which case I use the symbol <.

ADDITIONAL REFERENCES

Bosson, James E. (1964). *Modern Mongolian: a primer and reader*. Bloomington: Indiana University.
Lewis, M. B. (1947). *Teach yourself Malay*. London: English Universities Press.
Meiklejohn, Percy & Kathleen Meiklejohn (1958). Accentuation in Sarangani Manobo. In *Studies in Philippine linguistics*. Sydney: University of Sydney. 1–5.
Pater, Joe (1995). On the nonuniformity of weight-to-stress and stress preservation effects in English. Available as ROA-107 from the Rutgers Optimality Archive.
Shukla, Shaligram (1981). *Bhojpuri grammar*. Washington, D.C.: Georgetown University Press.
Street, John C. (1963). *Khalkha structure*. Bloomington: Indiana University.
Stuart, Don Graham (with the collaboration of Matthew M. Haltod) (1957). The phonology of the word in Modern Standard Mongolian. *Word* **13**. 65–99.
Tiwari, Udai Narain (1960). *The origin and development of Bhojpuri*. Calcutta: Asiatic Society.

The following codes are used in the 'note' columns in Tables I–V.

A No monosyllables.
B No light monosyllables.
C At most one heavy per word.
D At least one heavy per word.
E Rightmost even non-final syllable which is either heavy or followed by a (non-final) heavy.
F Pretonic heavies count as light.
G Light syllables occur only immediately following heavy syllables.
I Heavy syllables only occur initially.
X Not tail-canonically neighbourhood-distinct.
Y Not head-canonically neighbourhood-distinct.

	name	primary stress	secondary stress	note	FL	BL	FBL
single	Afrikaans	1L	none		4	4	4
	Abun	1R	none		4	4	4
	Diegueno (roots)	1R	none	B	4	4	4
	North Agul	2L	none		5	5	5
	Alawa	2R	none		5	5	5
	Mohawk	2R	none	A	5	5	5
	Cora	1L (2−), 3R (3+)	none		6	6	6
	Paamese	3R (3+), 1L (2−)	none	B,X	*	6	6
	Bhojpuri	3R (4+), 2R (3−)	not included	X	*	6	6
	Içuã Tupi	3R (5+), 2R (4−)	none	X,Y	*	*	*
	Bulgarian	lexical	none		4	4	4
dual	Gugu-Yalanji	1L	2R		6	6	6
	Sorbian	1L	none (3−), 2R (4+)	X	*	6	6
	Walmatjari	1L	2R or 3R (5+), 2R (4), none (3−)	Y	*	6	6
	Mingrelian	1L	3R (4+), none (3−)	X	*	*	*
	Armenian	1R	1L		5	5	5
	Udihe	1R	none (2−), 1L (3v)		5	6	6
	Anyula	2R	1L (4+), none (3−)		6	7	7
	Georgian	3R (3+), 2R (2−)	1L (5+), none (4−)		7	8	8

Table I
Quantity-insensitive single and dual patterns.

	name	primary stress	secondary stress	note	FL	BL	FBL
binary	Bagandji	1L	i2@1L		5	5	5
	Maranungku	1L	i2@1L	B	5	5	5
	Asmat	1R	i2@1R		5	5	5
	Araucanian	2L	i2@2L		6	6	6
	Anejom	2R	i2@2R		6	6	6
	Cavinena	2R	i2@2R	A	6	6	6
binary with lapse	Anguthimri	1L	i2@1L, no 1R		6	6	6
	Bidyara Gungabula	1L	i2@1L, no 1R	A	6	6	6
	Burum	1L	i2@1L, optional no 1R		5	5	5
	Garawa	1L	i2@2R, 1L, no 2L		6	6	6
	Indonesian	2R	i2@2R, 1L, no 2L (4+), none (3−)	X	*	8	8
	Piro	2R	i2@1L, 2R, no 3R		6	7	7
	Malakmalak	12@sL (3+), 1L (3−)	i2@2R (3+), none (3−)		6	6	6
binary with clash	Gosiute Shoshone	1L	i2@1L, 1R		c	6	6
	Tauya	1R	i2@1R, 1L		6	5	6
	Southern Paiute	2L (3+), 1L (2−)	i2@2L, 2R, no 1R (3+), none (2−)	B,Y	7	*	8
	Biangai	2R	i2@2R, 1L		7	6	7
	Central Alaskan Yupik	1R	i2@2L	B	6	6	6
ternary	Cayuvava	1L (2−), 3R (3+)	none (2−), i3@3R (3+)	A,X	*	8	9
	Ioway-Oto	2L	i3@2L		7	8	8

Table II
Quantity-insensitive binary and ternary patterns.

	name	primary stress	secondary stress	note	FL	BL	FBL
leftmost heavy, otherwise leftmost	Murik	12..89/1L	none	C	4	4	4
	Lithuanian	12..89/1L	none	D	4	4	4
	Amele	12..89/1L	none		4	5	5
	Khalkha Mongolian (Street)	12..89/1L	H		4	5	5
	Yidiŋ	12..89/1L	i2@m	B	5	*	5
	Kashmiri	12..78/12..78/1L	none		*	6	6
	Maori	12..89/12..89/1L	not included		5	5	5
	Khalkha Mongolian (Stuart)	12..89/2L	none		5	5	5
leftmost heavy, otherwise rightmost	Komi	12..89/9L	none		4	4	4
rightmost heavy, otherwise leftmost	Kuuku-Yau	12..89/9R	1L, H		5	5	5
	Dongolese Nubian	23..89/9R	H		5	5	5
	Khalkha Mongolian (Bosson)	23..891/9R	H		*	5	5
	Buriat	23..891/9R	1L, H		*	6	6
	Classical Arabic	1/23..89/ 9R	none		4	4	4
	Eastern Cheremis	23..89/9R	none		5	5	5
	Chuvash	12..89/9R	none		4	4	4

Table III

(continued on page 7)

	name	primary stress	secondary stress	note	FL	BL	FBL
rightmost heavy, otherwise rightmost	Golin	12..89/1R	none		5	4	5
	Meadow Cheremis	1/23..891/1R	none		5	4	5
	Mam	12..89/12/2R	none	B	5	5	5
	Klamath	12..89/23/3R	if 3R=SH, 2R=H then 2R		*	*	6
	Seneca	see note	i2@m<m	E	7	7	7
	Mountain Cheremis	23..89/2R	none		6	5	6
	Hindi (Jones)	23..891/2R	none		*	5	6
	Sindhi	23..891/2R	H		*	5	6
	Bhojpuri (Shukla/Tiwari)	23..891/2R	'Hm'H, m'LL, 1L	Y	*	*	6
	Hindi (Kelkar)	23..891/23..891/2R	H, i('LL)@m<m, m<i(LL')@m	X,Y	*	*	*

Table III
Quantity-sensitive unbounded patterns.

	name	primary stress	secondary stress	note	FL	BL	FBL
single	Maidu	12/2L	not included		4	4	4
	Hopi	12/2L	none	B	4	4	4
	English (verbs)	12/2R	not included		4	4	4
	Kawaiisu	12/2R	none	B	4	4	4
	Tümpisa Shoshone	21/1L	not included		5	5	5
	Javanese	21/1R	none		5	5	5
	Sarangani Manobo (Meiklejohn & Meiklejohn)	21/1R	none	B	5	5	5
	Awadhi	21/2R	none	B	5	5	5
	Malay (Lewis)	23/3R (3+), 12/2L (2−)	none		*	5	5
	Classical Latin	23/3R (3+), 1L (2−)	none	B	5	5	5
	Tiberian Hebrew	12/21/1R	not included		4	4	4
	English (nouns; Pater)	1@w3/234@sR	i('H,'LL)R		5	5	5
	Cairene Arabic	1@w3/23@sR	none	B	4	4	4
	Damascene Arabic	1@w3/23R	none		5	5	5
	Cyrenaican Bedouin Arabic	1@w3/23@sR (3+), 12/1R (2−)	i(H',LX')L (invs) (3+), none (2−)	B	*	*	*
	Hindi (Fairbanks)	12/2/34@sR (3+), 1L (2−)	i('H,'LL)R (invs) (3+), none (2−)	X	*	*	*
	Pirahã	123/123/ 123/123/1R	none	X	*	*	*

Table IV

(continued on page 9)

	name	primary stress	secondary stress	note	FL	BL	FBL
dual	Maithili	213/2R	1L	B	6	6	6
multiple	Cambodian	1R	H	B,G	5	5	5
	Yapese	12/1R	H		4	4	4
	Tongan	12/2R	H	B	4	4	4
	Sierra Miwok	12/2L	H	B	4	4	4
	Gurkhali	12/1L	m<H		4	4	4

Table IV
Quantity-sensitive bounded single, dual and multiple patterns.

	name	primary stress	secondary stress	note	FL	BL	FBL
binary	Western Aranda	12/2L (3+), 1L (2−)	i2@m, no 1R (3+), none (2−)		6	6	6
	Nyawaygi	12@sL	i('H,'LL)R		5	5	5
	Wargamay	12@sL	i('H,'LL)R, no 'H'L	B,I	6	6	6
	Berguener Romansh	12/2R	i('H,'LL)L		6	6	6
	Ancient Greek	12/2R	i('H,'LL)R		5	5	5
	Fijian	12/2R	i('H,'LL)R	B	5	5	5
	Romanian	12/2R	i2@m		5	5	5
	Seminole Creek	12@sR	i(H',LX')L	B	5	5	5
	Aklan	21/1R	i('H,'LL)@m<m		5	5	5
	Malecite/ Passamaquoddy	23@sR	i(H',LX')L		6	*	6
	Munsee	23@sR (3+), 12/2L (2−)	i(H',LX')L, no 1R (3+), none (2−)		*	6	6
	Cayuga	23@sR (3+), 1/0L (2−)	i(H',LX')L, no 1R (3+), none (2−)	B	6	6	6
	Manam	123/23/3R	i('H,'LL)@m<m		5	5	5
	Negev BedouinArabic	1@w3/23@sR (3+), 12/1R (2−)	i(H',LX')L (invs) (3+), none (2−)		*	*	*
	Bani-Hassan Arabic	1@w3/ 23@w2/2R	i('H,'LL)@m<m		5	5	5
	Palestinian Arabic	1/2/34@sR (3+), 1@w3/9R (2−)	i('H,'LL)L<m	B	*	*	*
	Ashéninca	234/324@s/324@sR	i('H,'LL)L<m (w2=H)	B,X	*	*	*
	Dutch	1@w4/23@sR	i('H,'LL)R		5	5	5
ternary	Estonian	1L	i('HX,'XLL, 'LL)L		6	6	6
	Hungarian	1L	i('HX,'XLL, 'LL)L, no 1R		6	6	6
	Sentani	12/2R	i('HX, 'XLX)@mL		7	6	7

Table V
Quantity-sensitive bounded binary and ternary patterns.

Appendix B: Proving convergence

This appendix provides a lemma and a theorem which are needed to establish convergence of the Forward and Backward learners.

Notation. π indicates a partition of some set of states Q. $B(q, \pi)$ denotes the block of states containing $q \in Q$ in some partition π. If π is a partition of a set S and S' is a subset of S, then the restriction of π to S' is the partition π' consisting of all those blocks B' that are non-empty and are the intersection of S' and some block of π. An acceptor A/π is the acceptor acquired by merging states in A which are in the same block in partition π. π_{nd} is the partition obtained by placing states with the same neighbourhood in the same block.

Lemma 1 Let S and S' be finite samples of L and denote the states of $PT(S)$ and $PT(S')$ with Q and Q' respectively. If

1. $S \subset S'$
2. $L = L(PT(S)/\pi)$
3. π' is a partition of $PT(S')$, and π is a restriction of π' to Q

then $L(PT(S)/\pi) \subseteq L(PT(S')/\pi')$.

Proof: Note that since $S \subset S'$, $Q \subset Q'$. Consider any

$$w = x_0x_1x_2 \ldots x_k \in L(PT(S)/\pi)$$

Then there is a path through $PT(S)/\pi$: $B(\lambda, \pi)$, $B(x_0, \pi)$, $B(x_0x_1, \pi)$, $B(x_0 \ldots x_k, \pi)$. Note that $B(\lambda, \pi)$ is the initial state and $B(x_0 \ldots x_k, \pi)$ is a final state in $PT(S)/\pi$ by definition of prefix tree and state-merging. Since π is a restriction of π' to Q, $B(\lambda, \pi)$, $B(x_0, \pi)$, $B(x_0x_1, \pi)$, $B(x_0 \ldots x_k, \pi)$ must also be a path in $PT(S')/\pi'$. Therefore $w \in L(PT(S')/\pi')$ and the lemma is proved.

Definition 1 We say that a finite sample S of a target language L is n-SATURATED if for all states up to depth n in $PT(S)$ there is no $w \in L$ such that $PT(S \cup \{w\})$ changes the neighbourhood of the states up to depth n.

Theorem 1 Let L be any regular language and let S and S' be finite samples of L such that

1. $S \subset S'$
2. $L = L(PT(S)/\pi_{nd})$
3. there is an n such that S is n-saturated

then $L(PT(S)/\pi_{nd}) \subseteq L(PT(S')/\pi_{nd})$.

Proof: Let Q and Q' denote the states of $PT(S)$ and $PT(S')$ respectively. Note that for any saturated states p, q in Q, $p, q \in Q'$, since $S \subset S'$. Since S is n-saturated, the neighbourhoods of p and q are the same in $PT(S)$ and $PT(S')$. Consequently,

$$B(p, \pi^S_{nd}) = B(q, \pi^S_{nd}) \text{ iff } B(p, \pi^{S'}_{nd}) = B(q, \pi^{S'}_{nd})$$

Thus π^S_{nd} is a restriction of $\pi^{S'}_{nd}$. Therefore, by Lemma 1, $L(PT(S)/\pi_{nd}) \subseteq L(PT(S')/\pi_{nd})$.

With Theorem 1 in place, it is simple to prove that the iterative versions of the Forward and Backward Learners can succeed, as explained in §8.